Text Summary Generation Techniques

Yashasvi Swapnesh Kumar Parikh, Narsingani Amisha Darshakbhai, Hetal Gaudani

Abstract: Pattern Recognition is pertinent field in autonomous text summarization for extraction of features from relative and non relative text documents. Here we provide empirical evidence that the method of Deep learning using RNN outperforms various techniques in terms of speed as well as metrics in abstractive summarization of multi-modal documents. We performed observational analysis on over 8 different techniques documented.

Keywords: Automatic Summarization, Natural Language Processing, Extractive Summary.

I. INTRODUCTION

Data Mining has seen rapid advances in recent years. This is particularly true for the case of text data as the advancement of software and hardware technologies have assisted the development of large data repositories. Many Documents are known for being lengthy. To our knowledge, some categories of documents contain duplicated and un useful information that does not require users attention. However, manually extracting non-duplicate information from documents requires considerable amount of effort. Automatic text summarization is the task of producing a concise and fluent summary while preserving key information content and overall meaning. It has two approaches. The approaches are Abstractive summarization and Extractive summarization. An extractive text summarization[14] means extraction of important information or sentences from the original document. Abstractive summarization[17][18] is a technique in which the summary is generated by generating novel sentences by either rephrasing the sentences or by introducing new words, instead of simply extracting and grouping the important sentences. Summaries and extracts provide a concise document description that reveal more than a document title, yet brief enough to be absorbed at a single glance. The need of summaries and extracts is increased by the large quantity of information currently available on-line. Traditional author-supplied indicative abstracts, when available, fulfill the need for a concise document description. The absence of author-supplied abstracts can be fulfilled with automatically generated summaries. [6]

II. RELATED WORK

J.N.Madhuri and Ganesh Kumar. R have employed Extractive Text Summarization Using Sentence Ranking [1]. C. Lakshmi Devasena and M. Hemalatha have implemented Automatic Text Categorization and Summarization using Rule Reduction[2]. Eliseo, Miriam and Mateus introduce a Text Mining Tool to Support Text Summarization[3]. Jingquiang Chen and HaiZhuge propose Extractive Summarization of Multi-Modal documents using RNN[4]. Lee JiHyung, Johana, Kyungmin Kim, Kim Nuri, Jaedong Lee have shown a summary generation using features of the document[5]. Julian, Jan, Francine, Daniel C. Brotsky and Steven B. Putz have used feature probabilities for text summarization[6]. Julian M. Kupiec, HinrichSchuetze have proposed a system for genre-specific summarization of text documents[8]. D.Ai, Y. Zheng, and D. Zhang has used latent semantic indexing to summarize text[21].

III. METHODOLOGY

A. Text Summarization Techniques

a. Extractive Text Summarization

The technique described by J.N.Madhuri and Ganesh Kumar.R[1] involves the selection of phrases and sentences from the source document to make up the new summary from the given text file or original document. A statistical method is used to perform an extractive text summarization on a single document[1].

The methodology used is as follows:

- Sentences are ranked by assigning weights and they are ranked based on their weights.[14][17]
- Highly ranked sentences are extracted from the input document
- This in turn directs to a high-quality summary of the input document
- Finally the generated summary is stored in the form of audio.

b. Rule Reduction Technique for Summarization

Rule reduction methodology by C. Lakshmi Devasena and M. Hemalatha[2] describes the use of classification to summarize the text document. The class labels are defined based on a set of examples of pre-classified documents used as a training set. This set of documents has been used to train the model. This method comprises an automatic text categorization and summarization approach to analyze the structure of input text. A text analyzer has been developed to derive the structure of the input text using rule reduction technique in three stages. The three stages of rule reduction technique are Token Creation,
Feature Identification and
Categorization, Summarization.

c. N-simple distance graph model

A text mining tool developed by Eliseo Reategui, Miriam Klemann and Mateus David Finco[3] is named as Sobec and is responsible for extracting graphs from texts. These graphs can be used in helping

Revised Manuscript Received on May 20, 2020.

* Correspondence Author

Yashasvi Swapnesh Kumar Parikh*, Computer Engineering, G. H. Patel College of Engineering & Technology, Vallabh Vidhyanagar, Anand, India. Email: parikhyashasvi@gmai.com

Narsingani Amisha Darshakbhai, Computer Engineering, G. H. Patel College of Engineering & Technology, Vallabh Vidhyanagar, Anand, India. Email: amishanarsingani@gmail.com

Hetal Gaudani, Computer Engineering, G. H. Patel College of Engineering & Technology, Vallabh Vidhyanagar, Anand, India. Email: hetalgaudani@gcet.ac.in

students to write summaries. The technique is based on the use of the graphs as a source of assisting students to further reflect on the main ideas of the text before getting to the actual task of writing. Thus, this tool provides an overview of the input text in the form of graph. A practical testing stated that the tool helped students reflect about the major ideas of the text and enhanced the writing of the summaries. Sobec has been developed using a particular mining algorithm. This algorithm was based on the n-simple distance graph model, in which nodes represent the main terms found in the text, and the edges represent adjacency information.

d. Recurrent neural network

Extractive Text-Image Summarization uses a neural-based multi-modal summarization method based on multi-modal RNN[4]. Here, Deep Learning techniques are found to be highly efficient in generating summaries. The approach is about summarizing documents with text and images by extracting sentences and images from the original document to generate the extractive multi model summary. Images play a vital role in delivering information to the audience[4]. Thus summarizing the images along with the text will increase the effectiveness of the summary generated. The model visits the input document twice. In the first visit, a bidirectional hierarchical RNN is used to encode the text document and the images. At the second visit, the model visits sentences one by one. Here the model uses a logistic classifier to calculate the summary probability of sentences. The features used for summary probability calculation are text coverage, text redundancy and image set coverage. Experiments state that this method outperforms the state-of-art neural summarization techniques.

e. Comparing word weighted values in contrast to various documents

Summary generation apparatus and method reflecting document feature[5], reflects the characteristics of a document in contrast to various documents. Sentences and words are extracted from multiple documents, in order to generate a candidate summary. The TextRank Algorithm calculates the importance of each sentence.[5] The method compares a word weighted value[5] of the original electronic documents to a word weighted value of the candidate summary - a word weighted value indicates whether a corresponding word is a core word or not in an individual document. It compares the similarities of the sentences[5] included in the candidate summary with the minimum score based on the TextRank Algorithm and a score calculation unit generating a final summary as output according to its convergence[5] with the final score of the previous summaries.

f. Feature probability calculation

Automatic method of extracting summarization using feature probabilities[6] is an iterative method. It firstly designates a sentence of the document, determines values for each feature of the sentence[6]. Further, it increases a score for the selected sentence based upon feature value and upon the probability associated with that value. A subset of the highest scoring sentences are extracted. The feature set here comprises a direct theme feature, a cue word feature[6] indicating whether the selected sentence summarizes the document, generating a thematic summary of the document. The method produces feature probabilities which automatically generates an extract for a machine readable representation of a natural language document using multiple features.

g. Latent Semantic Indexing

LSI - based text summarization method[21] exploits LSI to obtain the semantic structure of sentences and calculate the sentence similarities in the semantic space, and thus avoids the “bias” phenomenon caused by the word-based VSM(Vector Space Model)[23]. The LSI model projects the text from the observed high-dimensional surface space (composed of raw words) to the low-dimensional latent semantic space (composed of concepts)[21]. In doing so it eliminates the impact of orthogonal independence assumption in the VSM[22] developed by Deerwester, and determines the position of the text in the vector space more accurately.

h. Genre-specific summarization

System for Genre-specific summarization of documents[8] overcomes the problem of summarizing heterogeneous document collections by taking the genre, or type, of document into account when selecting summary sentences. A computer-implemented method for summarizing a document comprising: receiving a document having a plurality of document genres, determining a genre of document[[15] at the time of summarization, applying a genre-specific summarization routine[8] to selected sentence, evaluation being dependent at least in part on the genre of said document, assigning a summary score to the sentence based on the feature value produced by the summarization routine and selecting said sentence as a summary of the document if the summary score for the sentence is above a threshold value.

B. Proposed methodology

We have used Extractive summarization to summarize text documents. The first step was to split the text into sentences and sentences into words. All the content was formatted and converted into lowercase. All the punctuation marks were removed. Then after, words having fewer than 3 characters were removed. All the stopwords were removed.
The next step we did was Lemmatization.
Here, we changed the words in third person to first person and verbs in past and future tenses were changed to present tense. This step was followed by the reduction of words to their root form. All the stopwords that are present in the nltk_data directory, were removed from the content. Dataset used is the 20Newsgroup data set. This data set contains the news already grouped into key topics. There are 20 targets in the data set including some broad topics like Science, Politics, Sports, Religion, Technology. The next step after stemming and lemmatization is to calculate the frequency of words in the content. The final summary is generated by including the sentences that have the most frequent words in them.

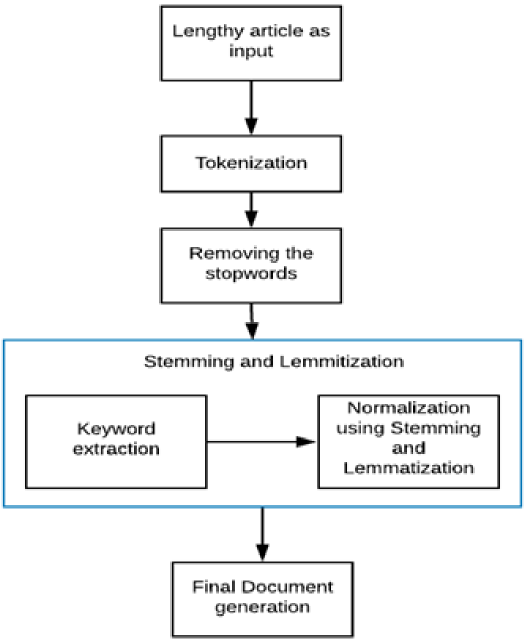
Lengthy article as input
Tokenization
Removing the stopwords
Stemming and Lemmitization
Keyword extraction
Normalization using Stemming and Lemmatization
Final Document generation

Figure I: Workflow

IV. EXPERIMENTS AND DISCUSSION

Table- I: Comparison of Text Summarization Techniques

Sr No.	Method Used	Observations
1	Extractive text summarization	Summary of a single text document can be generated using this algorithm based on the frequency of occurrence of words in the input text[1].
2	Rule reduction	This technique is used for Abstractive text Summarization to analyse the structure of text and generate the summary[2]. This technique is proven to generate more efficient summaries that Extractive Summarization.
3	Mining algorithm based on n-simple distance graph model	This methodology has been used to develop a text mining tool called Sobec [3]. This tool is used to generate the graphs that would assist the students in writing summaries [3].
4	Deep Learning Technique- neural-based extractive multi- modal summarization method based on multi- modal RNN [4].	Experiments show this method outperforms the state-of-the-art neural summarization methods[4]. This approach is proven to summarize the documents most efficiently.
5	Reflecting in contrast to various other documents using a TextRank Algorithm and TF-IDF method[23] (Term Frequency- Inverse Document Frequency)	Summary can be exported according to the importance of sentence. Also, in terms of existing technologies, the quality of the summary of output can also be improved.

6	Feature probabilities by reduced feature evaluation time. Direct Theme feature and a cue word feature.	The fraction of manual summary sentences faithfully reproduced for the El corpus 83%[6],the fraction of summary sentences correctly identified, can theoretically reach 100%. For the El corpus 42% of the document sentences extracted using the methods described matched a manual summary sentence [6].
7	Latent semantic indexing , Vector space model	The experimental results show that the summarization generated is of better quality than that generated by the traditional VSM[23] based approach, and thus validates its effectiveness. It evaluates 62%(approx)[21] of the documents accurately measured in terms of precision, recall and F-measure.
8	A genre- specific summarization routine. Classification methods - logistic regression, neural network, gradient descent algorithm.	The method produces a decision tree in which splits on the genre features are interleaved with splits on the summarization features may be used in the present invention.[8]

V. FUTURE SCOPE AND CONCLUSION

An automatically generated summary facilitates the Layman in getting a quick glance of long write-ups. Also the online filling forms procedure can be simplified. The word limit restrictions can be successfully achieved with efficient summary generation tools. This would reduce the human intervention in online form filling as the tool would be capable of minimizing the original uploaded content by abbreviating the number of words and sentences without changing the meaning of content. Moreover, the aged individuals who find difficulties in reading lengthy articles, indentures, forms, records, reports or any other quality paperwork can interpret the hard-coded text to get valuable knowledge by generating intelligible words. To add on to the advantages of Automated summarization, an extensible approach would be to create a job title by recommending a title job title could be recommended for a job opportunity based on the prerequisites specified by the recruiter in order to provide convenience to the job seekers or applicants. Moreover, Translators can utilize document summaries for obtaining a brief and a quick understanding of the original article. This would assist them in providing translated summary to the reviewer in comparatively shorter time span. Online summary generators can also be used in the presentation of documents in a manner that allows the user to quickly ascertain their contents by reviewing the documents obtained via electronic media which are often provided in such a volume that it is important to summarize them to get a brief summary rather than reading it completely.

REFERENCES

1. J.N.Madhuri and Ganesh Kumar.R , "Extractive Text Summarization Using Sentence Ranking"
2. C. Lakshmi Devasena and M. Hemalatha (2012) "Automatic Text Categorization and Summarization using Rule Reduction ", IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM -2012) March 30, 31, 2012
3. Eliseo Reategui, Miriam Klemann and Mateus David Finco, (2012) "Using a Text Mining Tool to Support Text Summarization", 12th IEEE International Conference on Advanced Learning Technologies.
4. Jingquiang Chen and HaiZhuge (2018) "Extractive Text-Image Summarization Using Multi-Modal RNN", 14th International Conference on Semantics, Knowledge and Grids (SKG)
5. JiHyungLee,,Johana, Kyungmin Kim, Kim Nuri, Jaedong Lee(2014), "Summary generation apparatus and method reflecting document feature"
6. Julian M. Kupiec,Jan O. Pedersen,Francine R. Chen,Daniel C. Brotsky,Steven B. Putz(1995), "Automatic method of extracting summarization using feature probabilities"
7. InderjeetMani,EugenioCiurana,NicholasD'Aloisio-Montilla,Bart K. Swanson(2012) "Method and apparatus for automatically summarizing the contents of electronic documents", Proceedings of Workshop on Text Summarization.
8. Julian M. Kupiec,HinrichSchuetze (1999) "System for genre-specific summarization of documents"
9. Leonid Batchilo, Valery Tsourikov, Igor Sovpel(2007) "Computer Based Summarization Of Natural Language Documents"
10. Goldstein et al(1999)., "Summarizing Text Documents: Sentence Selection and Evaluation Metrics", Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.
11. Kenton M. Lyons, Barbara Rosario, Trevor Pering, Roy Want(2012), "Methods and systems to summarize a source text as a function of contextual information"
12. Wei JiaCai, Xiao XiaoLian, Shixia Liu, Shimei Pan, Wei Hong Qian, Yang Qiu Song, Qiang Zhang. Michelle Xue Zhou(2009), "Producing a visual summarization of text documents"
13. Amjad Abu-Jbara and DragomirRadev. (2011), " Coherent citation-based summarization of scientific papers.", In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 500–509.
14. Harold P Edmundson(1969). "New methods in automatic extracting", Journal of the ACM (JACM) 16, 2 (1969), 264–285.
15. Kessler et al.(1997), " Automatic Detection of Text Genre" , Proceedings of ACL 35 and EACL 8, Morgan Kaufmann Publishers, San Francisco, California, pp. 32–38.
16. Abuobieda, A., Salim, N., Albaham, A. T., Osman, A. H., & Kumar, Y. J. (2012), "Text summarization features selection method using pseudo genetic-based model. In International conference on information retrieval knowledge management (pp.193–197).
17. G Erkan and Dragomir R. Radev(2004), "LexRank: Graph-based Centrality as Salience in Text

Summarization", Journal of Artificial Intelligence Research, Research, Vol. 22, pp. 457-479.

18. Udo Hahn and Martin Romacker (2001), "The SYNDIKATE text Knowledge base generator", Proceedings of the first International conference on Human language technology research, Association for Computational Linguistics , ACM, Morristown, NJ, USA.
19. Balahur, Alexandra.,Lloret, Elena., Boldrini, Ester., Montoyo, Andrés., Palomar, Manuel., &Martı´nez-Barco, Patricio (2009)," Summarizing threads in blogs using opinion polarity.", Proceedings of the workshop on events in emerging text types (pp. 23–31).
20. Gong, Y.H., Liu, X(2002), "Generic text summarization using relevance measures and latent semantic analysis.", Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 19–25. ACM, New York
21. D. Ai, Y. Zheng, and D. Zhang(2010), "Automatic text summarization based on latent semantic indexing", Journal of Artificial Life and Robotics, Springer, vol. 15, no. 1
22. Deerwester S, Dumais S, Furnas G, et al (1990), "Indexing by latent semantic analysis", J Am Soc Inform Sci 41(6):391–407
23. Christian S. Perone(2011), "Short introduction to Vector Space Model (VSM)", blogs-christianperone, Machine Learning :: Text feature extraction (tf-idf) – Part I
24. Saranyamol C S and Sindhu L(2014), "A Survey on Automatic Text Summarization", International Journal of Computer Science and Information Technologies, Vol. 5(6)
25. Chengqing Z (2008),"Statistical natural language processing (in Chinese)", Tsinghua University Press, Beijing
26. Hearst, MA (1997), "TextTiling: segmenting text into multi-paragraph subtopic passages", Comput Linguistics 32(1)
27. PadmaPriya, G. and K. Duraiswamy(2014), "An approach for text summarization Using deep learning algorithm", ISSN: 1549-3636.
28. G. Salton and C. Buckley(1998), "Term-weighting approaches in automatic text retrieval," Information Processing & Management, vol. 25, no. 5, pp. 513-523
29. Ronald A. Fein, et al(1999), "Document summarizer for word processors"

AUTHORS PROFILE

Yashasvi Swapnesh Kumar Parikh is pursuing her Bachelor's in Computer Engineering at G. H. Patel College of Engineering & Technology (GCET) at VallabhVidyanagar, Anand, Gujarat, India. She is a final year student of Computer Engineering. She will be graduating in May 2020. She has been the Tech-Lead and Treasurer in the Core Committee of the Computer Society of India (CSI) at the Student Branch of GCET. She is functioning as the Mentor of Computer Society of India at the Student Branch of the Institute.

Narsingani Amisha Darshakbhaiis a final year Bachelor's student in Computer Engineering, graduating in May 2020 from G. H. Patel College of Engineering & Technology (GCET), VallabhVidyanagar, Anand, Gujarat, India. She has been a professional Android Developer since 2018 and has served as Web Developer Head at Indian Society of Technical Education (ISTE) Student Branch of GCET. She has developed face recognition and detection systems, Language translator apps for social purpose.

Prof Hetal Gaudani obtained Master's in Computer Engineering from Gujarat Technological University and Bachelor in Computer Engineering from Dharmsinh Desai University. Her main research interest includes Big Data Analytics, Machine Learning and Computer Vision. She has published many research papers in conferences and reputed journals. She has developed online IVR, GSM and biometric Based Hostel Management System for GCET Girls Hostel. She was Member of the Board of Studies in Computer Engineering in July 2011 in Sardar Patel University.